Hierarchical Clustering of Shotgun Proteomics Data

Ville R. Koskinen, Patrick A. Emery, David M. Creasy, and John S. Cottrell‡

A new result report for Mascot search results is described. A greedy set cover algorithm is used to create a minimal set of proteins, which is then grouped into families on the basis of shared peptide matches. Protein families with multiple members are represented by dendrograms, generated by hierarchical clustering using the score of the nonshared peptide matches as a distance metric. The peptide matches to the proteins in a family can be compared side by side to assess the experimental evidence for each protein. If the evidence for a particular family member is considered inadequate, the dendrogram can be cut to reduce the number of distinct family members. ***Molecular & Cellular Proteomics 10: 10.1074/mcp.M110.003822, 1–12, 2011.***

In shotgun proteomics, a mixture of proteins, which may be as complex as a whole cell lysate, is digested to peptides prior to chromatographic separation and analysis by mass spectrometry. Database searching of the tandem MS (MS/MS)[1] spectra delivers matches to peptide sequences. Using these matches to deduce which proteins were present in the original sample is surprisingly difficult because many of the peptide sequences in a typical search result can be assigned to more than one protein.

A comprehensive description of the "Protein Inference Problem" can be found in the review by Nesvizhskii and Aebersold (1). More recently, computational tools for protein inference and estimation of protein false discovery rate (FDR) have been reviewed by Li *et al.* (2), who observed that they can be categorized as deterministic approaches (DBParser (3), Mass Sieve (4), EPIR (5), Isoform Resolver (6), DTASelect (7), ProteinScape (8), IDPicker (9, 10), PROVALT (11)) or probabilistic approaches (Qscore (12), PRISM (13), ProteinProphet (14), PRO_PROBE (15), PANORAMICS (16), and EBP (17)). A review by Shi and Wu (18) contains additional discussion of how peptide uniqueness and detectability can be used, such as proteotypic peptides (19) and a Bayesian model that penalizes a protein for the absence of a match to an expected peptide (20).

Other approaches include protein interaction network information as a basis for accepting protein identifications that might otherwise be rejected as unsafe, such as proteins identified by a single peptide (21); spectral networks, in which overlapping uninterpreted MS/MS spectra are combined into longer chains, then mapped directly to protein sequences (22); the classification of peptides according to a fully characterized gene model (23, 24); and MAYU analysis to estimate the FDR for an existing set of protein identifications (25).

Fig. 1 illustrates the fundamental ways in which proteins can be related through shared matches. In this discussion, we assume that most individual peptide matches are reliable. That is, some type of threshold has been applied to eliminate the bulk of random matches, resulting in a known and acceptable FDR.

Shared peptide matches are mainly the result of sequence redundancy among the database entries. Causes include

- Proteins that are alternative splice forms of the same gene
- Products of related genes (gene paralogs)
- Conserved regions and motifs common to many proteins
- Multiple entries for the same protein with sequencing or typographical errors
- Multiple entries for the same protein with polymorphisms
- Homologous proteins from related organisms

The extent of sequence redundancy in the public databases varies considerably (26). At one extreme, with very high redundancy, are the comprehensive, nonidentical databases such as National Center for Biotechnology Information (NCBI) nr and UniRef100. At the other extreme, curated, nonredundant databases such as Swiss-Prot and International Protein Index. Sequence redundancy is compounded by identification ambiguity. Matching is a statistical process and, however stringent the scoring and filtering, there will be some level of incorrect identifications. Random matches are not such a problem because they are more or less uniformly distributed across the database entries. It is systematic mis-identification because of ambiguous mass values that most affects protein inference. Depending on the mass spectrometer type and accuracy, it may be difficult or impossible to distinguish between I and L, Q and K, F and oxidized M, E or D and de-amidated Q or N, etc.

A report created from database search results can take a maximal or minimal approach to listing the identified proteins.

From Matrix Science Ltd., London, UK
Received July 29, 2010, and in revised form, February 22, 2011
Published, MCP Papers in Press, March 29, 2011, DOI 10.1074/mcp.M110.003822

[1] The abbreviations used are: MS/MS, tandem MS; BLAST, Basic Local Alignment Search Tool; FDR, false discovery rate; iTRAQ, Isobaric Tags for Relative or Absolute Quantitation.

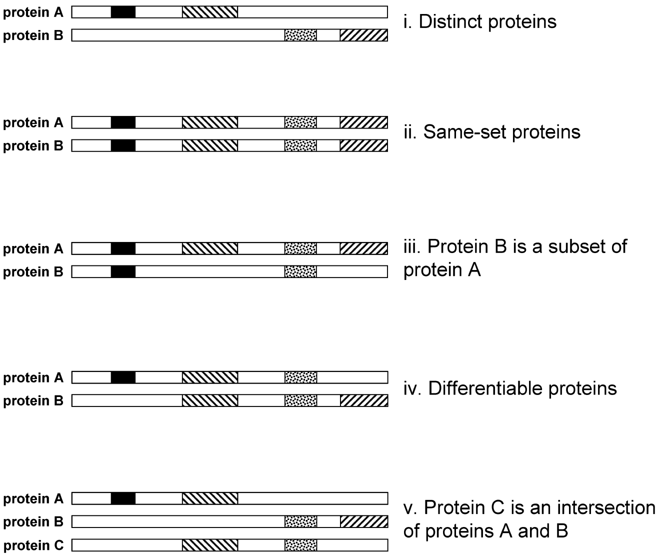


FIG. 1. **(*i*) A and B are distinct proteins, with no shared matches.** There is evidence for both and both should be listed in the report; (*ii*) A and B are same-set proteins (also termed indistinguishable (1) or equivalent (3, 9)). The report should make it clear that both are equally valid assignments of the peptide matches and that either could be present in the sample. The possibility that both are present is rejected by parsimony; (*iii*) B is a subset of protein A. B may be present in the sample, but there is no evidence for this, so by parsimony it is relegated to an inferior status or dropped entirely from the report; (*iv*) A and B are related through shared matches but there is evidence for both being present in the sample. They should both be listed in the report, ideally with some indication of their relationship; (*v*) Protein C is an intersection protein, a subset of the combined matches to A and B (also termed subsumable (1, 3, 9)). By parsimony, it is relegated to an inferior status or dropped entirely from the report.

The maximal list is all database entries that contain one or more of the identified peptides. The minimal list is the smallest set of database entries that accounts for all the identified peptides. The mechanism for selecting a minimal list is often described as Occam's razor or the principle of parsimony.

The maximal approach is only useful for the very smallest of searches, where manual inspection or evidence from other sources will be used to decide which proteins were truly present in the sample. Publishing a maximal list is discouraged by journal guidelines because of the risk that a researcher using the data may not understand that the sample actually contained only a fraction of the proteins listed, and there are many possible ways to select a shorter list of proteins that would account for all the observed peptides. This does not mean that a minimal list is an end in itself, or the ideal representation of the search results for all purposes (27). For example, it is perfectly possible for the protein with the largest number of high scoring peptide matches or the greatest sequence coverage to be dropped, as illustrated by the hypothetical case in Fig. 2.

Computing a protein score or probability from the scores or probabilities of the assigned peptide matches is a popular way of ordering a report, so that the proteins with the largest number of strong peptide matches rise to the top. It is arguable whether protein scores have any deeper meaning. In a large scale experiment, where the minimal protein list may contain hundreds or even thousands of proteins, strict ordering by protein score can cause similar proteins to become widely scattered, making relationships, such as the presence of isoforms, difficult to discern.

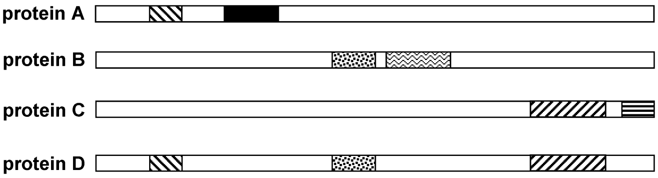


FIG. 2. **A truly minimal list of proteins would contain only A, B, and C.** Protein D would be an intersection protein, even though it might have the greatest coverage and the highest score.

Of course, the true goal of a shotgun proteomics experiment is not the creation of a table of proteins for a publication; it is to gain insight into a biological system. Because there is no immediate prospect of replacing the researcher's biological knowledge with any kind of expert system, the key requirement is to present the search results in the clearest possible manner, making it easy to 'drill down' and inspect the evidence for proteins of interest. The report should facilitate answering questions such as:

- For which proteins do we need to make antibodies?
- Is there evidence for a particular isoform of this protein?
- Does this protein carry a biologically interesting modification or polymorphism?
- Which proteins have been up- or down-regulated?

Grouping identified proteins into families based on sequence homology dramatically simplifies the interpretation of a result report because it makes it easier to locate the proteins of interest. All against all alignment using sequence homology is computationally intensive, so it is usually implemented by searching a preclustered database, such as ProteinCenter (Proxeon A/S, Odense, Denmark) or by using a prebuilt index that maps database entries into families, such as UniGene (28). Mascot Integra (Matrix Science Ltd., London, UK) is a proteomics data management system that implements real-time Basic Local Alignment Search Tool (BLAST)-based clustering of the proteins found in a database search result, but this is time consuming when a large number of proteins are selected unless a powerful BLAST server is available.

We have found that clustering by means of shared peptide matches is an acceptable surrogate for homology based clustering, and have extended this by performing hierarchical clustering of each protein family using the score of the non-shared peptide matches as a distance metric. Hierarchical clustering has been applied previously to matrix assisted laser dissociation ionization-MS mass values in connection with Peptide Mass Fingerprinting (29) but not, to our knowledge, for grouping proteins based on shared peptide matches. A new report is described, in which each protein family is represented by a dendrogram (or cladogram), illustrating whether family members are closely or distantly related. A greedy set-cover algorithm is used to converge rapidly on a minimal list of proteins. The peptide matches to the proteins in a family can be compared side by side to assess the experimental evidence for each of them. If the evidence for a family member is considered inadequate, the dendrogram can be cut to reduce the number of distinct family members.

EXPERIMENTAL PROCEDURES

Searches of a public domain data set distributed for the Association of Biomolecular Resource Facilities iPRG2008 study (30) are used to illustrate points in the discussion. Proteins from a mouse liver differential expression experiment (details not provided) were digested with trypsin, alkylated with methyl methanethiosulfonate, labeled with an isobaric tag for relative and absolute quantitation (iTRAQ) tag, combined, and then separated into 13 fractions by strong cation exchange chromatography. The fractions were analyzed on an Applied Biosystems 3200 QTRAP system (AB Sciex, Foster City, CA) producing 29 raw files and 41,977 spectra. Peak lists were generated by iPRG committee members in a variety of formats. The Mascot Generic Format peak list set used here was downloaded from https://www.abrf.org/index.cfm/group.show/ProteomicsInformaticsResearchGroup.53.htm (archive password iprgcode).

Searches were performed using Mascot 2.3 (Matrix Science Ltd., London, UK). Automatic decoy mode was used, which generates and searches a separate database of random sequences in which the number of entries and the length of each entry is the same as in the target database. Search parameters were:

Enzyme : Trypsin/P

Fixed modifications : iTRAQ4plex (K),iTRAQ4plex (N-term),Methylthio (C)

Variable modifications : Acetyl (Protein N-term),Gln->pyro-Glu (N-term Q),Oxidation (M)

Mass values : Monoisotopic

Peptide Mass Tolerance : ± 0.9 Da

Fragment Mass Tolerance : ± 0.6 Da

Max Missed Cleavages : 1

Instrument type : ESI-TRAP

Number of queries : 33,191

Modification names and compositions are taken from Unimod (http://www.unimod.org). The mass tolerances may seem high for a QTRAP, but they were set by inspection of preliminary search results and were necessary to accommodate the observed errors. Several different sequence databases were searched, as described in the Results & Discussion section.

The algorithm used to create the protein family report, which is part of Mascot 2.3, is summarized in Fig. 3. The report is generated by a Perl script that calls the Mascot Parser library to read data from the Mascot result file.

RESULTS AND DISCUSSION

In the new report, proteins are grouped into a family if they have significant matches to one or more distinct peptide sequences in common. Matches with scores below the significance threshold play no part in the grouping because they have an unacceptable chance of being random and should not be used as evidence to link or differentiate proteins.

If two proteins have the same set of peptide matches, the distance between them is zero. If they have a mix of shared and nonshared matches, the distance between them is the sum of the score excesses over threshold of all the distinct, nonshared matches in one protein, because discarding these would make one protein a subset of the other, based on the shared matches. In this calculation, each distinct peptide sequence is represented once by the match with the highest score, irrespective of charge state or modifications state. Note that this distance measure is asymmetric, and the score distance to make protein A into a subset of protein B will not be the same as that to make B a subset of A. The smaller distance is always chosen.

There are some subtleties to this procedure. Consider the case of two proteins which have different peptide matches to the same spectrum with the same score. Only one of these matches can be correct, but we do not know which. (Unless the spectrum contains fragments from multiple peptides, as discussed below.) An example is where the two peptide sequences differ only in exchange of I and L. These sequences may behave differently in biological terms but, if the scores are the same, there is simply no evidence from the mass spectrometry data to distinguish the two possibilities (31, 32). Such a match should make no contribution to the distance between the two proteins.

Now, consider the case in which we have two proteins with different peptide matches to the same spectrum, but the scores are not the same. Assume the score threshold is 40 and one match has a score of 60 and the other has a score of 70. Again, only one of these matches can be correct; but it is not the same as if they were independent matches to different queries. Extending the logic that matches to the same spectrum with the same score correspond to a distance of zero, matches to the same spectrum with different scores contribute a distance that is the score difference. In this example, the distance would be 10. If the two matches came from different queries, and could be treated independently, the distance contribution would be either $(70 - 40) = 30$ or $(60 - 40) = 20$, depending on which one had to be sacrificed to make one protein into a subset of the other.

This neglects the possibility that matches to two different peptides are obtained because the spectrum is a mixed one, containing fragments from a pair of isobaric and co-eluting peptides. In most cases, the scores from mixed spectra are poor, and it would be unusual to get significant matches to both components. Should this occur, and one or both sequences are not represented independently with a higher score elsewhere in the search results, from a different time point or in a different modification state, then the distance among the proteins will be under-represented. In the limiting case of equal scores, it will be as if there was no match to the mixed spectrum.

To create the dendrogram, we first compute a distance matrix, which is the distance between each pair of proteins. The two proteins separated by the smallest distance are joined to create a node, and the length of the branches from

```
1. From the search results, create an initial list of proteins, ordered by protein score (note 1)
2. Take the highest scoring protein on the list
3. Find all other proteins in the same family:
    3.1. select all peptide matches with homology score or better
    3.2. for each peptide match, select all the proteins that contain this match and remove from
         the initial list (note 2)
    3.3. for each new protein, select all new peptide matches with homology score or better
    3.4. loop until no further proteins or peptide matches remain
4. For each protein in the family, make a list of the distinct peptide sequences. That is, ignoring
   differences in modification state and precursor charge. Where there are duplicate matches to
   a sequence, the representative score for the sequence is the highest one
5. Using this set of distinct peptide sequences, divide and group the proteins into same-set
   proteins and subset proteins, which includes intersections (note 3).
    5.1. Same-set proteins are collapsed into a single family member (note 4)
    5.2. Proteins that are subsets, including intersections, are relegated to secondary status
    5.3. Perform hierarchical clustering of the family members (note 5)
6. Loop from step 2 until no more peptide matches remain with homology score or better
```

FIG. 3. **The grouping and filtering algorithm used in the new report.**

Note 1. The protein score is the average score threshold plus the sum over all peptide matches of the excess of the peptide match score over the score threshold. Nonsignificant peptide match scores make no contribution to the protein score, which is essential to avoid scores from random matches accumulating into substantial protein scores when the search contains a large number of spectra relative to the number of database entries. If a protein contains a single significant peptide match, the protein score is the same as the peptide match score. If there are duplicate matches, each contributes to the score. The default peptide match score threshold is the Mascot homology threshold, which is an empirical estimate of whether the score is an outlier. The significance threshold is 5% by default, and can be changed in the user interface.

Note 2. When selecting the proteins that contain a given match, if there are other matches to the same spectrum with identical scores, proteins containing these other matches are also selected. The rationale is that, if there is no score difference between two sequences, we cannot distinguish them and they should be treated symmetrically. An example would be two peptides that had identical sequences apart from interchanges of I and L. Where there are duplicate matches, it can happen that for one spectrum, two similar sequences get the same score whereas, for another spectrum, they get different scores. In such cases, the sequences are treated as distinguishable.

Note 3. Finding all possible intersections so as to achieve maximum parsimony is an “NP-hard” problem. We use an iterative method to rapidly find a solution that is acceptably close to the optimum. The algorithm is based on the “greedy set cover algorithm” (35) used in IDPicker (9). We have added two pruning steps that further reduce the number of proteins to inspect. In the following pseudo code, a free peptide means a peptide that is not contained by any protein in the result set S1:

```
1. Let P be the set of all proteins in the family and S1 and S2 be empty sets of proteins.
2. While there are proteins in P:
   2.1. Select a protein p from P such that p covers the most free peptides, meaning p has the maximum number of peptides not yet in any protein in S1.
   2.2. If at least one of p ’s peptides is contained by a protein in S1:
      2.2.1. Let Q be a subset of S1 where all proteins in Q share at least one peptide with p.
      2.2.2. For each protein q in Q: if all of q ’s peptides are contained by p plus the other proteins in Q, q would be an intersection after the addition of p. Move q from S1 to S2.
   2.3. Move p from P to S1.
   2.4. For each protein q in P: move q from P to S2 if q is an intersection in S1, meaning all of q ’s peptides are contained by some set of proteins in S1.
3. The set of proteins S1 contains a heuristic minimum set of proteins covering all peptides in this family, whereas S2 contains proteins that are subsets or intersections of proteins in S1. (The reason step 2.2 is before step 2.3 is that this makes it easier to prove S1 never contains proteins that are subsets or intersections.)
```

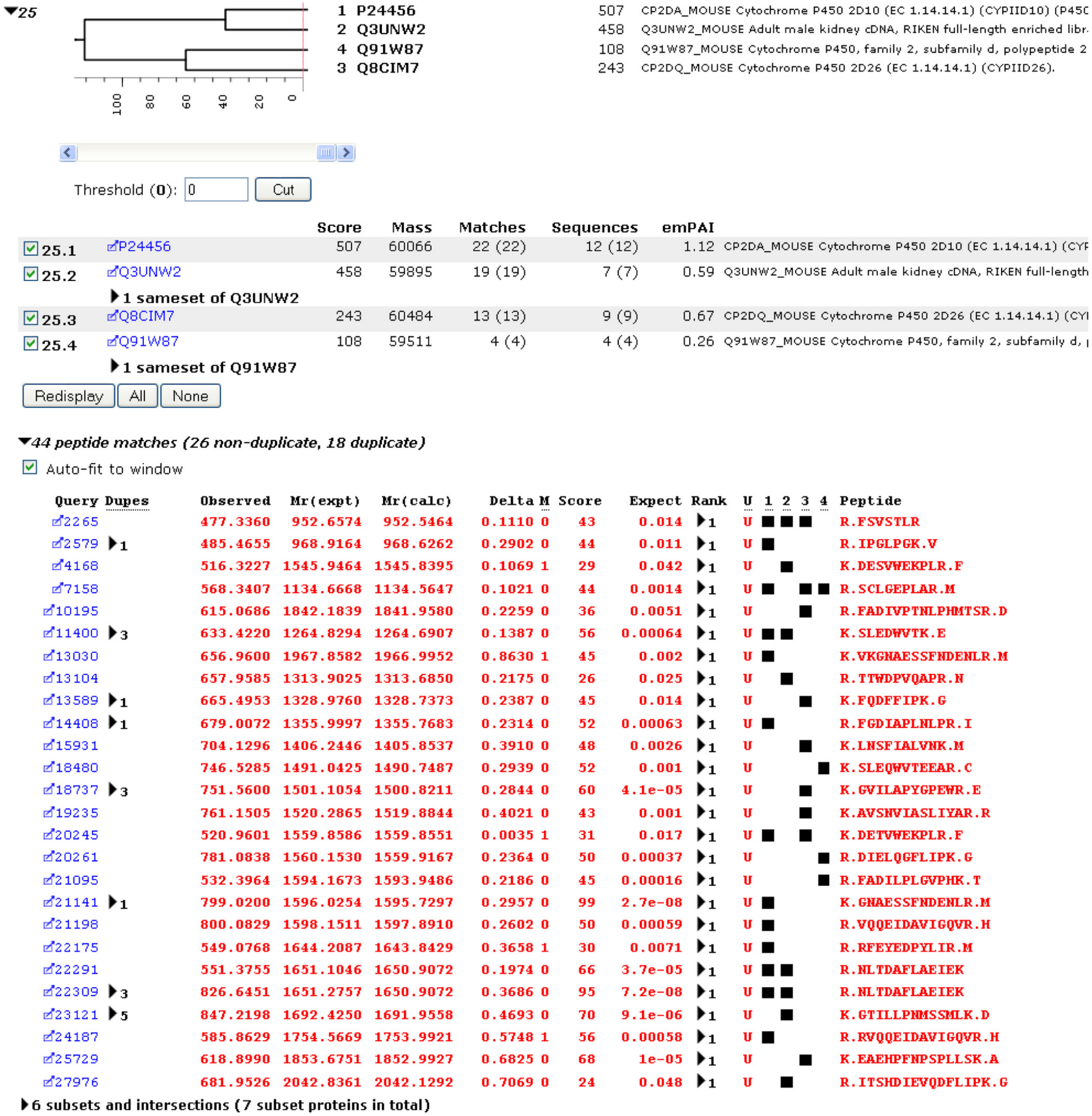
Note 4. In our terminology, a family is a set of proteins related by shared peptide matches. A family member is a set of proteins corresponding to the same set of peptide matches. There is no evidence to distinguish the proteins in a family member and no reason to prefer one over another. The choice of one protein from a family member as the anchor, to be listed first and used to label the dendrogram, does not indicate a preference.

Note 5. The distance metric is the score of the nonshared peptide sequences, but similar results would be obtained using peptide lengths or simply the count. The reason for choosing score is that our confidence in the match increases with score. We contend that a nonshared match close to the score threshold is less evidence for the presence of two proteins than a match with a very high score. Functions from the Cluster 3.0 library (Michiel de Hoon, University of Tokyo, Human Genome Center) are called to perform agglomerative, single linkage clustering.

the node is the score distance among the proteins. The two joined proteins are removed from the list, and the distances between the new node and all other remaining proteins (or nodes) computed. The process is repeated until only one node remains. When the dendrogram (or tree) is drawn, the order is chosen to avoid any branches crossing. There is no other significance to the order of the branches, and there are many possible ways to order the branches so as to avoid crossings. In the tabular part of the report, proteins are sorted in order of decreasing score, which will often be different from the order of the dendrogram branches.

It is common for two members of a larger family to have no shared matches. Every protein in the family is linked to others by means of shared matches, or they would never have been grouped together, but this doesn’t mean that there are going to be shared matches between every pair of family members. More unusually, a family member will appear to have no shared matches with any other member. This can happen when all the proteins that link the member into the family are relegated to subset status as intersection proteins. Fig. 2 illustrates such a case.

Fig. 4 shows one family from searching the iPRG2008 data against the IPRG2008 study database (53,826 mouse proteins plus 74 contaminants). When the report first loads, only the dendrogram is displayed for each family, labeled with accessions, scores, and descriptions for the anchor proteins for each family member. (An anchor protein is a representative of the same-set proteins selected in some consistent manner, *e.g.*



1	P24456	507	CP2DA_MOUSE Cytochrome P450 2D10 (EC 1.14.14.1) (CYPIID10) (P450
2	Q3UNW2	458	Q3UNW2_MOUSE Adult male kidney cDNA, RIKEN full-length enriched libr.
4	Q91W87	108	Q91W87_MOUSE Cytochrome P450, family 2, subfamily d, polypeptide 2
3	Q8CIM7	243	CP2DQ_MOUSE Cytochrome P450 2D26 (EC 1.14.14.1) (CYPIID26).

		Score	Mass	Matches	Sequences	emPAI	
☑ 25.1	P24456	507	60066	22 (22)	12 (12)	1.12	CP2DA_MOUSE Cytochrome P450 2D10 (EC 1.14.14.1) (CYF
☑ 25.2	Q3UNW2	458	59895	19 (19)	7 (7)	0.59	Q3UNW2_MOUSE Adult male kidney cDNA, RIKEN full-length
	▶1 sameset of Q3UNW2						
☑ 25.3	Q8CIM7	243	60484	13 (13)	9 (9)	0.67	CP2DQ_MOUSE Cytochrome P450 2D26 (EC 1.14.14.1) (CYI
☑ 25.4	Q91W87	108	59511	4 (4)	4 (4)	0.26	Q91W87_MOUSE Cytochrome P450, family 2, subfamily d,
	▶1 sameset of Q91W87						

▼44 peptide matches (26 non-duplicate, 18 duplicate)

☑ Auto-fit to window

Query	Dupes	Observed	Mr(expt)	Mr(calc)	Delta	M	Score	Expect	Rank	U	1	2	3	4	Peptide
2265		477.3360	952.6574	952.5464	0.1110	0	43	0.014	▶1	U	■	■	■		R.FSVSTLR
2579	▶1	485.4655	968.9164	968.6262	0.2902	0	44	0.011	▶1	U	■				R.IPGLPGK.V
4168		516.3227	1545.9464	1545.8395	0.1069	1	29	0.042	▶1	U		■			K.DESVWEKPLR.F
7158		568.3407	1134.6668	1134.5647	0.1021	0	44	0.0014	▶1	U	■		■	■	R.SCLGEPLAR.M
10195		615.0686	1842.1839	1841.9580	0.2259	0	36	0.0051	▶1	U			■		R.FADIVPTNLPHMTSR.D
11400	▶3	633.4220	1264.8294	1264.6907	0.1387	0	56	0.00064	▶1	U	■	■			K.SLEDWVTK.E
13030		656.9600	1967.8582	1966.9952	0.8630	1	45	0.002	▶1	U	■				K.VKGNAESSFNDENLR.M
13104		657.9585	1313.9025	1313.6850	0.2175	0	26	0.025	▶1	U		■			R.TTWDPVQAPR.N
13589	▶1	665.4953	1328.9760	1328.7373	0.2387	0	45	0.014	▶1	U			■		K.FQDFFIPK.G
14408	▶1	679.0072	1355.9997	1355.7683	0.2314	0	52	0.00063	▶1	U	■				R.FGDIAPLNLPR.I
15931		704.1296	1406.2446	1405.8537	0.3910	0	48	0.0026	▶1	U			■		K.LNSFIALVNK.M
18480		746.5285	1491.0425	1490.7487	0.2939	0	52	0.001	▶1	U				■	K.SLEQWVTEEAR.C
18737	▶3	751.5600	1501.1054	1500.8211	0.2844	0	60	4.1e-05	▶1	U			■		K.GVILAPYGPEWR.E
19235		761.1505	1520.2865	1519.8844	0.4021	0	43	0.001	▶1	U			■		K.AVSNVIASLIYAR.R
20245		520.9601	1559.8586	1559.8551	0.0035	1	31	0.017	▶1	U	■		■		K.DETVWEKPLR.F
20261		781.0838	1560.1530	1559.9167	0.2364	0	50	0.00037	▶1	U				■	R.DIELQGFLIPK.G
21095		532.3964	1594.1673	1593.9486	0.2186	0	45	0.00016	▶1	U				■	R.FADILPLGVPHK.T
21141	▶1	799.0200	1596.0254	1595.7297	0.2957	0	99	2.7e-08	▶1	U	■				K.GNAESSFNDENLR.M
21198		800.0829	1598.1511	1597.8910	0.2602	0	50	0.00059	▶1	U	■				R.VQQEIDAVIGQVR.H
22175		549.0768	1644.2087	1643.8429	0.3658	1	30	0.0071	▶1	U	■				R.RFEYEDPYLIR.M
22291		551.3755	1651.1046	1650.9072	0.1974	0	66	3.7e-05	▶1	U	■	■			R.NLTDAFLAEIEK
22309	▶3	826.6451	1651.2757	1650.9072	0.3686	0	95	7.2e-08	▶1	U	■	■			R.NLTDAFLAEIEK
23121	▶5	847.2198	1692.4250	1691.9558	0.4693	0	70	9.1e-06	▶1	U		■			K.GTILLPNMSSMLK.D
24187		585.8629	1754.5669	1753.9921	0.5748	1	56	0.00058	▶1	U	■				R.RVQQEIDAVIGQVR.H
25729		618.8990	1853.6751	1852.9927	0.6825	0	68	1e-05	▶1	U			■		K.EAEHPFNPSPLLSK.A
27976		681.9526	2042.8361	2042.1292	0.7069	0	24	0.048	▶1	U		■			R.ITSHDIEVQDFLIPK.G

▶6 subsets and intersections (7 subset proteins in total)

FIG. 4. **A protein family found by searching the iPRG2008 data against the IPRG2008 study database.** The display has been partly expanded to show details of the anchor proteins and the peptide matches. Only peptide matches with scores above a 5% homology threshold are displayed.

first in an alphabetical sort of accession strings.) This allows the person viewing the report to scan rapidly down a long list of proteins, looking for families of interest. The report is paged, with a default of 10 families per page, but this can be changed to display a larger number or all families in a single list.

If a family is of interest, the details can be displayed by clicking on the small triangle to expand that section of the report. It is clear from the protein descriptions that family 25 contains four cytochrome P450 proteins. Fig. 4 shows the appearance when this family is partially expanded. A list of the anchor proteins with additional information and expansion buttons for any same-sets proteins is followed by a table of the peptide matches. To simplify the display for publication, the table has been filtered so that only peptide matches with scores above a 5% significance threshold are displayed. Controls for such filters are at the head of the report.

For each of the anchor proteins, beside the protein score and mass, there are counts of the number of matches and the number of distinct sequences. In each column, the first number is the total count whereas the number in parentheses is the count for matches above the significance threshold. Because we have filtered the report to only include matches above this threshold, the two numbers are always the same in the figure. Anchor proteins are in order of protein score, and it can be seen that family member 25.1 (P24456 cytochrome P450 2D10) is represented by significant matches to 12 distinct sequences compared with family member 25.4 (Q91W87 cytochrome P450 2D22), which has only four. A very approximate estimate of the relative abundance of each protein is provided by Exponentially Modified Protein Abundance Index values (33).

Most of the information in the peptide match table will be self explanatory. Where there are duplicate matches to the same sequence with the same precursor charge and modification state, only the highest scoring match is shown by default. Clicking on the small triangle in the Dup[licat]es column will expand the table to show the other matches in-line. Similarly, up to the ten highest scoring matches per MS/MS spectrum are saved in the result file, but only the match assigned to the protein is displayed by default. Clicking on the small triangle in the Rank column will expand the table to show the other matches in-line. The column headed U contains a U if the peptide match is unique to the family. Because each family is grouped on matches above the significance threshold, only matches below this threshold can be nonunique, and such matches have been filtered out in the figure.

The adjacent columns provide a mapping between the family members and the peptide matches. A square marker at the intersection of a row and column indicates that the peptide match is found in the family member. Usually, the interest will be in making a pair-wise comparison, and too many columns can confuse the eye, so checkboxes above the table can be used to select a smaller set of proteins. Peptide match rows are only displayed if the match is found in at least one of the selected proteins, so the number of rows will also decrease and the table becomes much easier to comprehend.

Selecting family members 25.1 (P24456 cytochrome P450 2D10) and 25.2 (Q3UNW2 cytochrome P450 2D9) would immediately show that they have three distinct sequences in common, and that we would have to discard matches to four distinct sequences to make 25.2 into a subset of 25.1. The score excess over threshold for these four sequence totals 43, which corresponds to the distance on the dendrogram from the origin to the point at which these two family members are joined. Cutting the dendrogram at (say) 45 would make Q3UNW2 into a subset of P24456 and reduce the family to three members.

A decision about whether to include Q3UNW2 in a list of identified proteins comes down to whether one is willing to treat these nine matches to four distinct sequences as unreliable. Looking at the peptide match table, three of these sequences have weak matches, barely above threshold. Most of the score comes from GTILLPNMSSMLK, with six matches, the highest score of 70 corresponding to an expect value of 9.1E-6. In such a case, expanding the row to show the alternative matches to the spectrum is important because it could be that there was another match with a similar score that belonged to a different family. If so, our confidence in using this match as evidence for Q3UNW2 might be reduced. But, this is not the case. A further step might be to run a BLAST search, and see whether a very similar peptide is found in one of the other family members, raising the possibility that the matched peptide contains a polymorphism. This can be done by clicking on two hyperlinks. In a search of Swiss-Prot 2010_04 using BLAST 2.2.23 (NCBI, Bethesda, MD), the next best match is to the sequence we are currently comparing it to: P24456 cytochrome P450 2D10

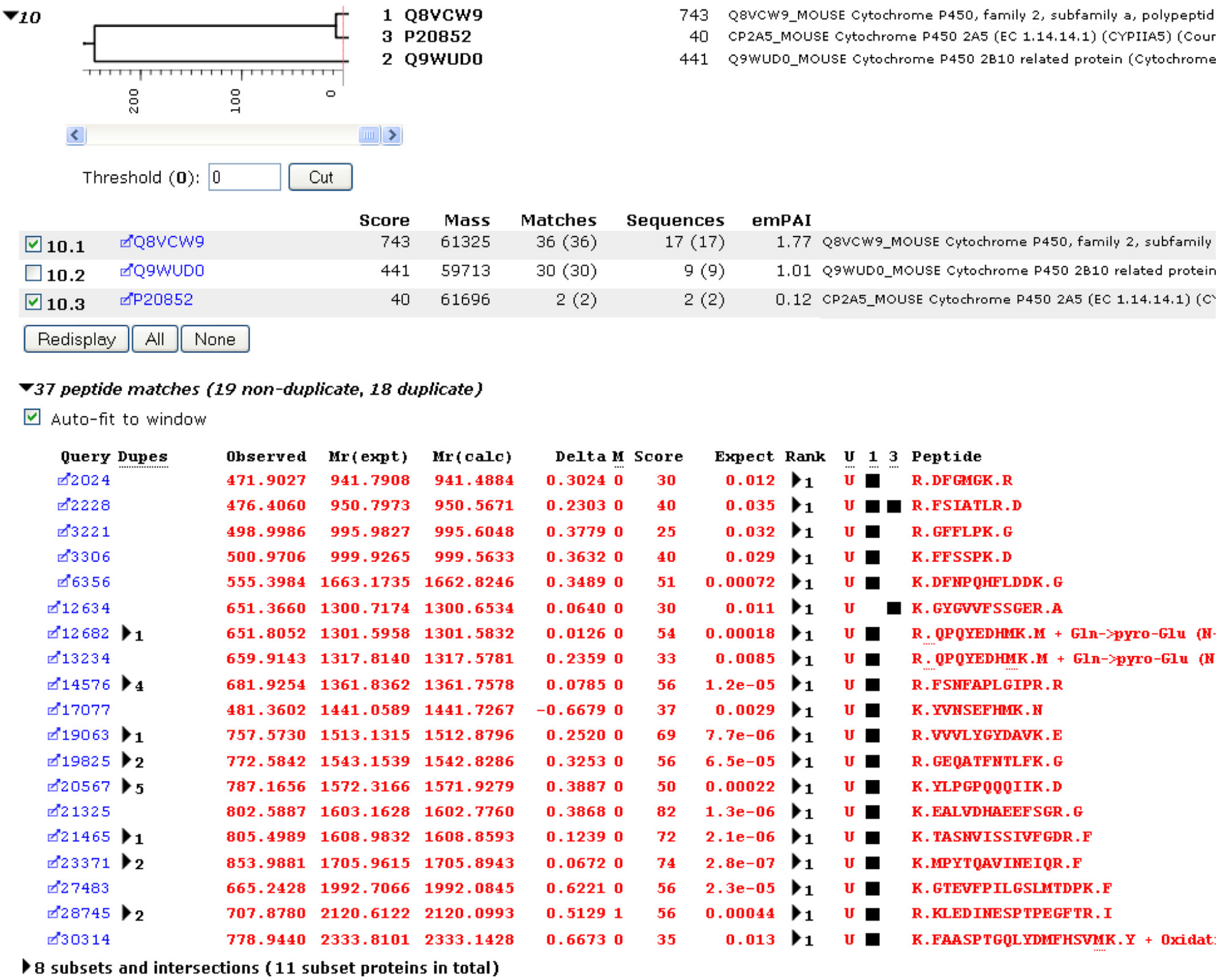
```
Query  1    GTILLPNMSSMLK  13
            G+IL+PNMSS+LK
Sbjct  395  GSILIPNMSSVLK  407
```

No known mutants at these positions are listed in either Swiss-Prot entry, and this is a strong match so, on balance, one might choose to accept it as sufficient evidence that the sample contained cytochrome P450 2D9.

Slightly higher in the report, there is another family of cytochrome P450 proteins, illustrated in Fig. 5. Note the large difference in protein score between family member 10.1, score 743, and member 10.3, score 40. In a report sorted by protein score, these two proteins would be widely separated. When grouped into a family, the two members are separated by a score difference of seven, corresponding to a single peptide, GYGVVFSSGER. If it was not for this match, with a score of 30 and an expect value of 0.011, family member 10.3, (P20852 cytochrome P450 2A5), represented by just two distinct sequences, would become a subset of 10.1 (Q8VCW9 cytochrome P450 2A512). Although it is a subjective decision, the evidence for the presence of P20852 is slim, and many would choose to drop it.

In a very long list of proteins, such decisions need to be reduced to simple rules that can be automated in software. These cases illustrate how the dendrogram can be used as a short-cut to making such decisions. By cutting all dendrograms at a score difference of 10 or 20 or whatever value is preferred, family members for which there is only tenuous evidence can be relegated to subset status automatically.

The iPRG2008 study was designed to benchmark the quality of protein inference on a realistically complex data set, and the results were released as a poster that analyzed the numbers of true and false positive proteins reported by the participants (https://www.abrf.org/ResearchGroups/ProteomicsInformatics



1	Q8VCW9	743	Q8VCW9_MOUSE Cytochrome P450, family 2, subfamily a, polypeptid
3	P20852	40	CP2A5_MOUSE Cytochrome P450 2A5 (EC 1.14.14.1) (CYPIIA5) (Cour
2	Q9WUD0	441	Q9WUD0_MOUSE Cytochrome P450 2B10 related protein (Cytochrome

Threshold (0): 0

		Score	Mass	Matches	Sequences	emPAI	
☑ 10.1	Q8VCW9	743	61325	36 (36)	17 (17)	1.77	Q8VCW9_MOUSE Cytochrome P450, family 2, subfamily
☐ 10.2	Q9WUD0	441	59713	30 (30)	9 (9)	1.01	Q9WUD0_MOUSE Cytochrome P450 2B10 related protein
☑ 10.3	P20852	40	61696	2 (2)	2 (2)	0.12	CP2A5_MOUSE Cytochrome P450 2A5 (EC 1.14.14.1) (C

▼*37 peptide matches (19 non-duplicate, 18 duplicate)*

☑ Auto-fit to window

Query	Dupes	Observed	Mr(expt)	Mr(calc)	Delta	M	Score	Expect	Rank	U	1	3	Peptide
2024		471.9027	941.7908	941.4884	0.3024	0	30	0.012	▶1	U	■		R.DFGMGK.R
2228		476.4060	950.7973	950.5671	0.2303	0	40	0.035	▶1	U	■	■	R.FSIATLR.D
3221		498.9986	995.9827	995.6048	0.3779	0	25	0.032	▶1	U	■		R.GFFLPK.G
3306		500.9706	999.9265	999.5633	0.3632	0	40	0.029	▶1	U	■		K.FFSSPK.D
6356		555.3984	1663.1735	1662.8246	0.3489	0	51	0.00072	▶1	U	■		K.DFNPQHFLDDK.G
12634		651.3660	1300.7174	1300.6534	0.0640	0	30	0.011	▶1	U		■	K.GYGVVFSSGER.A
12682	▶1	651.8052	1301.5958	1301.5832	0.0126	0	54	0.00018	▶1	U	■		R.QPQYEDHMK.M + Gln->pyro-Glu (N-
13234		659.9143	1317.8140	1317.5781	0.2359	0	33	0.0085	▶1	U	■		R.QPQYEDHMK.M + Gln->pyro-Glu (N
14576	▶4	681.9254	1361.8362	1361.7578	0.0785	0	56	1.2e-05	▶1	U	■		R.FSNFAPLGIPR.R
17077		481.3602	1441.0589	1441.7267	-0.6679	0	37	0.0029	▶1	U	■		K.YVNSEFHMK.N
19063	▶1	757.5730	1513.1315	1512.8796	0.2520	0	69	7.7e-06	▶1	U	■		R.VVVLYGYDAVK.E
19825	▶2	772.5842	1543.1539	1542.8286	0.3253	0	56	6.5e-05	▶1	U	■		R.GEQATFNTLFK.G
20567	▶5	787.1656	1572.3166	1571.9279	0.3887	0	50	0.00022	▶1	U	■		K.YLPGPQQQIIK.D
21325		802.5887	1603.1628	1602.7760	0.3868	0	82	1.3e-06	▶1	U	■		K.EALVDHAEEFSGR.G
21465	▶1	805.4989	1608.9832	1608.8593	0.1239	0	72	2.1e-06	▶1	U	■		K.TASNVISSIVFGDR.F
23371	▶2	853.9881	1705.9615	1705.8943	0.0672	0	74	2.8e-07	▶1	U	■		K.MPYTQAVINEIQR.F
27483		665.2428	1992.7066	1992.0845	0.6221	0	56	2.3e-05	▶1	U	■		K.GTEVFPILGSLMTDPK.F
28745	▶2	707.8780	2120.6122	2120.0993	0.5129	1	56	0.00044	▶1	U	■		R.KLEDINESPTPEGFTR.I
30314		778.9440	2333.8101	2333.1428	0.6673	0	35	0.013	▶1	U	■		K.FAASPTGQLYDMFHSVMK.Y + Oxidat

▶8 subsets and intersections (11 subset proteins in total)

FIG. 5. **A family of cytochrome P450 proteins from the same report as Fig. 4.**

ResearchGroup/EPosters/iPRG2008_InitialResultsPoster.pdf). The reference list was a consensus of the results obtained by the iPRG committee members, with groups of proteins based on shared matches clustered using UniRef50. An alignment between the iPRG2008 study results and the list of proteins obtained from the Mascot family report is available as supplementary material. At a peptide FDR of 5%, as determined by the decoy database search, there are zero decoy proteins with significant matches to two or more distinct peptide sequences. In the target database, after cutting the dendrograms at a score of 10, 825 protein accessions are clustered into 191 families containing 227 family members. Relative to the reference list, this corresponds to 219 true positive and 4 false positive proteins. Two participants in the study reported 254 class 1,2,3 true positive proteins for the same number of false positives, whereas nine others reported more true positives at the expense of a greater number of false positives. The reference list does not detail how individual accessions have been assigned to "detectable isoforms," so it is not possible to say precisely which family members might be considered the false positives.

One way to test how well an algorithm approaches an ideally parsimonious list of proteins is to compare the results for the same search run against databases of different size, in which the larger ones are super-sets of the smaller. To test this with the iPRG2008 data, we took the mouse sequences from Swiss-Prot by applying a filter of reviewed: yes AND taxonomy:10090 to UniProt release 2010_04 (http://www.uniprot.org/). Both canonical and isoform sequences were downloaded in Fasta format. We then widened the taxonomy in four further steps until we were taking all the entries in Swiss-Prot.

If these five Fasta files were to be searched directly, the results would be distorted by the variation in significance threshold resulting from the change in the number of entries. That is, a match that was just above the significance threshold

TABLE I

The number of protein hits for different reports as a function of target database size. Taxonomy was widened in four steps from mouse to all entries in Swiss-Prot. The total size of the Fasta file was kept constant by reversing and appending the balance of the Swiss-Prot entries. (i) Peptide Summary with no filters. (ii) Peptide Summary after discarding peptide matches with scores below the significance threshold and proteins that do not contain at least one match that is 'bold red'. (iii) Family members in the new report. (iv) Families in the new report

Taxonomy ID	10090	9989	314146	33154	1
Taxonomy	Mus musculus	Rodentia	Euarchontoglires	Fungi/Metazoa	All
Selected entries	23781	34355	77125	149726	545039
Reversed entries	521258	510684	467914	395313	0
cRAP entries	112	112	112	112	112
Total entries	545151	545151	545151	545151	545151
(i) Peptide Summary, no filters	189	240	286	421	484
(ii) Peptide Summary, filtered	171	180	183	184	177
(iii) Family members	172	182	185	184	180
(iv) Families	155	156	155	151	150

in one of the smaller databases might be lost in the search of a larger database. To correct for this, we reversed and appended all the sequences in Swiss-Prot that were not part of the selected taxonomy, making the size of the database invariant for all searches.

Because the sample was known from earlier searches to contain bovine trypsin, the individual Swiss-Prot databases were searched in combination with a contaminants database, cRAP (http://www.thegpm.org/crap/index.html), to ensure that this was not a source of variation between searches. Mascot automatic decoy mode simulates a search of a separate, randomized, decoy database, enabling the calculated significance threshold to be adjusted to achieve a verified FDR (for peptide matches) between 4.9% and 5.1%. Examination of the decoy matches showed that no decoy protein had significant matches to more than one distinct peptide sequence. By applying a rule that a protein must contain significant matches to at least two distinct peptide sequences, we expect that few, if any, false proteins are reported.

Table I lists the number of reported proteins as a function of the size of the target database. The Peptide Summary report has been part of Mascot since version 1.5. It uses the principal of parsimony, but with two limitations. First, nonshared matches with scores well below the significance threshold can prevent one protein from becoming a subset of another. Second, the algorithm does not attempt to find intersection proteins. The number of proteins increases from 189 in the mouse database to 484 in the full Swiss-Prot. (The standard Mascot Peptide Summary report uses the identity threshold whereas the Family Summary uses the homology threshold, but for comparison purposes, all counts in Table I are based on matches above the homology threshold.) The Peptide summary more closely approaches a minimal list of proteins if two filters are applied. Peptide matches that are below a 5% significance threshold are discarded and a protein is only listed if it contains at least one match that is "bold red." That is, a match that is the highest scoring for the spectrum (red) and where a match to the spectrum has not appeared in any higher scoring protein (bold). This gives a much smaller list of proteins, which changes little with database size (range 171 to 184). The 'require bold red' filter removes the majority of the intersection proteins and proteins that would be subset proteins except for random matches, but misses some because it depends on the order in which the proteins appear in the report. The final two rows of Table I are for family members in the new report and for families. The difference between the filtered peptide summary and count of family members is small, indicating that there are few intersection proteins in this particular result, if any. In fact, the filtered peptide summary tends to contain one or two fewer proteins. This is caused by the "require bold red" filter removing proteins that it should not. For example, in the search of the mouse database, CP238_MOUSE is listed as family member because of a significant match to EALIDHGEEFSGR. This is not the top-ranked match to the spectrum, and there are no other "bold red" matches for this protein, so it gets dropped from the filtered Peptide Summary. The count of families is relatively flat with database size (range 150 to 156). It drops slightly for the two larger databases because homologous proteins from other organisms occasionally act as bridges to connect families that had no shared matches in proteins from the narrower taxonomy.

To compare clustering by shared peptide matches with clustering by sequence homology, we aligned the family report for a search of a mouse EST database with the results of the same search after mapping the original accessions into UniGene cluster accessions. UniGene (28) is a system for automatically partitioning GenBank sequences, including ESTs, into a nonredundant set of gene-oriented clusters. The sequence database was the Mus division of EST sequences from European Molecular Biology Laboratory (EMBL) release 104, containing 4,852,146 sequences. The UniGene index was Mus musculus Build #182.

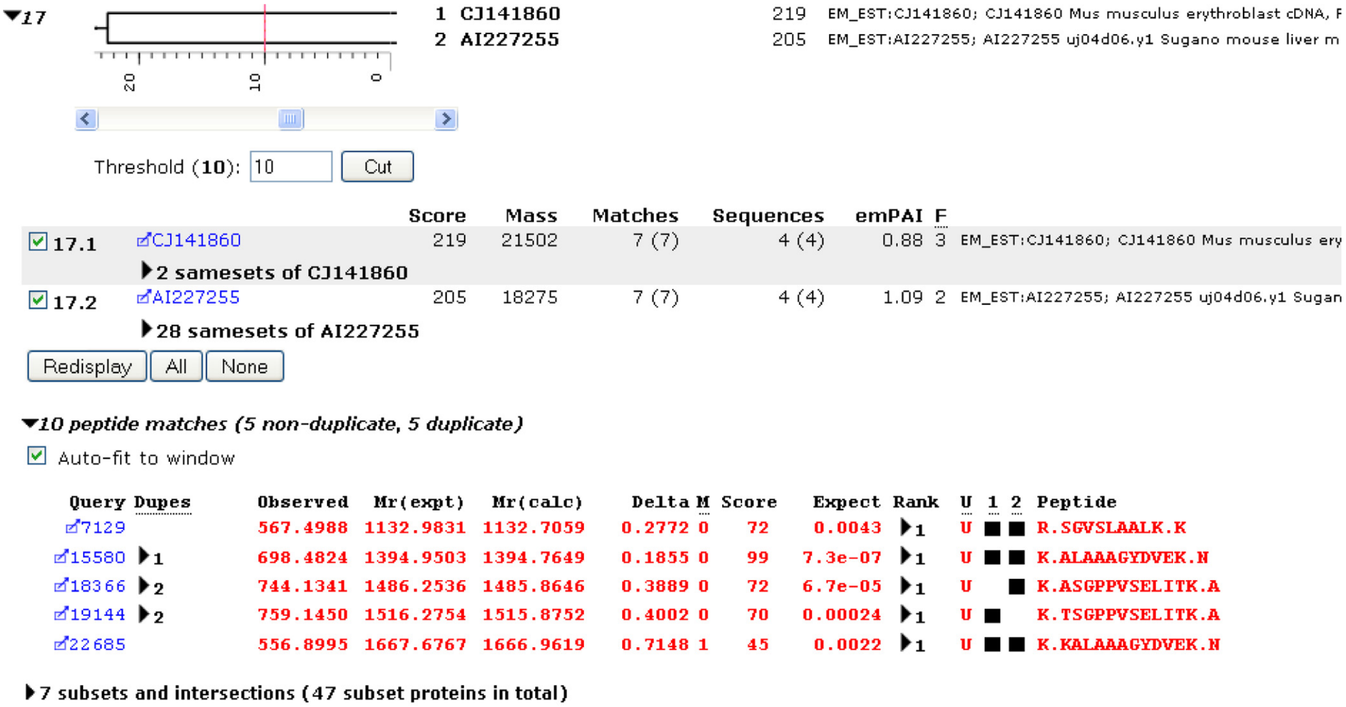
Table II shows the alignment for the first 20 families in the EST report when the significance threshold is set to give a 5% FDR for peptide matches and all dendrograms are cut at a score of 10. If the two clustering methods have similar out-

TABLE II
Alignment of EMBL and UniGene accessions for the first 20 families in the results from searching the iPRG2008 data against Mus division of EST sequences from EMBL release 104

Family	EST Accession	UniGene Cluster	UniGene Description
1.1	BY012418	Mm.31018	Cyb5 Cytochrome b-5
1.2	W91084	Mm.31018	Cyb5 Cytochrome b-5
2.1	AA002359	Mm.14796	Mgst1 Microsomal glutathione S-transferase 1
3.1	CX120581	Mm.289810	Rpl14 Ribosomal protein L14
4.1	BI145268	Mm.15537	Cyp1a2 Cytochrome P450, family 1, subfamily a, polypeptide 2
4.2	BI221323	Mm.15537	Cyp1a2 Cytochrome P450, family 1, subfamily a, polypeptide 2
5.1	AW012478	Mm.20764	Cyp2c29 Cytochrome P450, family 2, subfamily c, polypeptide 29
5.2	AI526761	Mm.38963	Cyp2c50 Cytochrome P450, family 2, subfamily c, polypeptide 50
5.3	AA238951	Mm.379575	Cyp2c54 Cytochrome P450, family 2, subfamily c, polypeptide 54
5.4	AI047293	Mm.20764	Cyp2c29 Cytochrome P450, family 2, subfamily c, polypeptide 29
5.5	AI132230	Mm.379575	Cyp2c54 Cytochrome P450, family 2, subfamily c, polypeptide 54
6.1	BI327647	Mm.6696	Rdh7 Retinol dehydrogenase 7
6.2	BI218937	Mm.6696	Rdh7 Retinol dehydrogenase 7
7.1	AW413050	Mm.332844	Cyp3a11 Cytochrome P450, family 3, subfamily a, polypeptide 11
8.1	AA000970	Mm.328601	Transcribed locus, strongly similar to 60S ribosomal protein L7a
9.1	CK023210	Mm.330160	Hspa5 Heat shock protein 5
9.2	AA065715	Mm.330160	Hspa5 Heat shock protein 5
9.3	BG861518	Mm.330160	Hspa5 Heat shock protein 5
9.4	AU036073	Mm.412745	Transcribed locus, strongly similar to heat shock cognate 71 kDa protein
10.1	CX232350	Mm.16660	P4hb Prolyl 4-hydroxylase, beta polypeptide
10.2	BE307099	Mm.16660	P4hb Prolyl 4-hydroxylase, beta polypeptide
10.3	BF119796	Mm.16660	P4hb Prolyl 4-hydroxylase, beta polypeptide
11.1	BI145775	Mm.174372	Cyp2d22 Cytochrome P450, family 2, subfamily d, polypeptide 22
12.1	BG865446	Mm.292803	Ces3 Carboxylesterase 3
12.2	AA647338	Mm.88078	Es1 Esterase 1
12.3	AI116604	Mm.292803	Ces3 Carboxylesterase 3
13.1	BI330877	Mm.26741	Ugt2b1 UDP glucuronosyltransferase 2 family, polypeptide B1
13.2	AI097842	Mm.291575	Ugt2b5 UDP glucuronosyltransferase 2 family, polypeptide B5
13.3	AI118428	Mm.291575	Ugt2b5 UDP glucuronosyltransferase 2 family, polypeptide B5
13.4	AA511527	Mm.300095	Ugt1a@ UDP glycosyltransferase 1 family, polypeptide A cluster
14.1	BF011466	Mm.371545	Rplp0 Ribosomal protein, large, P0
15.1	BI147765	Mm.295534	Es31 Esterase 31
16.1	W12409	Mm.380435	Rplp2 Ribosomal protein, large P2
16.2	AV482679	Mm.380435	Rplp2 Ribosomal protein, large P2
16.3	BY412823	Mm.380435	Rplp2 Ribosomal protein, large P2
17.1	CJ141860	Mm.170587	Hist1h1e Histone cluster 1, H1e
17.2	AI227255	Mm.193539	Hist1h1c Histone cluster 1, H1c
18.1	BI408452	Mm.222825	Pdia6 Protein disulfide isomerase associated 6
19.1	AA690820	Mm.319719	Rpl13 Ribosomal protein L13
20.1	BU054756	Mm.22560	Cyb5r3 Cytochrome b5 reductase 3

comes, we would expect the members of each family to map to the same UniGene cluster. Ten of the families have a single member, five map to a single UniGene accession and five map to multiple accessions (families 5, 9, 12, 13, and 17). The question is whether this indicates overclustering, where one or two shared matches join proteins that have little homology otherwise.

The upper part of Fig. 6 shows the matches for family 17. In UniGene, CJ141860 is assigned to Mm.170587 (Histone cluster 1, H1e) whereas AI227255 is assigned to Mm.193539 (Histone cluster 1, H1c). The alignment of the two sequences using ClustalW (34) is shown in the lower part of Fig. 6, with the matches highlighted in red. The region in which we have matches is near identical, differing by only a single residue. The rest of the sequences have limited homology, but there are no peptide matches in these regions, so the distance between the sequences in the dendrogram is small. It is not easy to say whether we have evidence for two proteins. There is just one pair of nonshared peptide matches, and these only differ by a single residue. On the other hand, the scores for the matches are good, with expect values of the order of 10^{-4}, there are three matches to each sequence, and the two sequences are in perfect alignment, which need not be the case if one match was random. There is the possibility that both matches are to the same primary sequence but one or both are modified close to the N terminus in a way that creates a mass difference of 30 Da. There is no single entry in Unimod that could account for this, but there are possibilities if multiple modifications are considered. If this question was important, the way to resolve it would be to acquire more



▼17

1	CJ141860	219	EM_EST:CJ141860; CJ141860 Mus musculus erythroblast cDNA, F
2	AI227255	205	EM_EST:AI227255; AI227255 uj04d06.y1 Sugano mouse liver m

Threshold (**10**): 10 Cut

		Score	Mass	Matches	Sequences	emPAI	F	
☑ 17.1	CJ141860	219	21502	7 (7)	4 (4)	0.88	3	EM_EST:CJ141860; CJ141860 Mus musculus ery
	▶2 samesets of CJ141860							
☑ 17.2	AI227255	205	18275	7 (7)	4 (4)	1.09	2	EM_EST:AI227255; AI227255 uj04d06.y1 Sugan
	▶28 samesets of AI227255							

Redisplay All None

▼*10 peptide matches (5 non-duplicate, 5 duplicate)*

☑ Auto-fit to window

Query	Dupes	Observed	Mr(expt)	Mr(calc)	Delta	M	Score	Expect	Rank	U	1	2	Peptide
7129		567.4988	1132.9831	1132.7059	0.2772	0	72	0.0043	▶1	U	■	■	R.SGVSLAALK.K
15580	▶1	698.4824	1394.9503	1394.7649	0.1855	0	99	7.3e-07	▶1	U	■	■	K.ALAAAGYDVEK.N
18366	▶2	744.1341	1486.2536	1485.8646	0.3889	0	72	6.7e-05	▶1	U		■	K.ASGPPVSELITK.A
19144	▶2	759.1450	1516.2754	1515.8752	0.4002	0	70	0.00024	▶1	U	■		K.TSGPPVSELITK.A
22685		556.8995	1667.6767	1666.9619	0.7148	1	45	0.0022	▶1	U	■	■	K.KALAAAGYDVEK.N

▶7 subsets and intersections (47 subset proteins in total)

```
CLUSTAL 2.0.12 multiple sequence alignment


AI227255        ----------XFLTSXILIMSEAAPAAPAAAPPAEKAPAKKKAAKKPAGVRRKASGPPVS 50
CJ141860        SPGQSLCFRLEFSLLTRFAMSETAPAAPAAPAPAEKTPVKKKARKAAGGAKRKTSGPPVS 60
                           *     : ***:*******..****:*.**** * ..*.:**:******

AI227255        ELITKAVAASKERSGVSLAALKKALAAAGYDVEKNISXIKLGPEEPDEQGHPVANQWHRC 110
CJ141860        ELITKAVAASKERSGVSLAALKKALAAAGYDVEKNNSRIKLGLKSLVSKGTLVQTKGTGA 120
                *********************************** * **** :.  .:*  * .:   .

AI227255        LRLLQTQQEGAVLARPNPRLRRQARTNAIEACGSX--PR-SPNMATCC------------ 155
CJ141860        SGSFKLNKK-AASGEAKPKAKRAGAAKAKKPAGAAKKPKKAAGTATAKKSTKKTPKKAKK 179
                   :: ::: *. ...:*: :* . ::* :..*:   *: :.. **.

AI227255        -------
CJ141860        PAAAAGA 186
```

FIG. 6. **(*upper*) Family 15 from the results of searching the iPRG2008 data against the Mus division of EST sequences from EMBL release 104.** (*lower*) Alignment of the two sequences using ClustalW.

data in a targeted experiment. Otherwise, clustering these two proteins together seems like a reasonable way to present the results, even though UniGene places the two sequences into different gene families.

Inspection of families 5, 9, 12 and 13 reveals a similar story. Clustering on the basis of shared peptide matches sometimes groups proteins that belong to different UniGene clusters, but it doesn't appear to group unrelated proteins. For example, the Cytochrome P450 proteins are cleanly divided into family numbers 4 (Cyp1a), 5 (Cyp2c), 7 (Cyp3a), and 11 (Cyp2d). It is to be expected that shared matches will be less discriminating than sequence homology alignment, which attempts to align complete protein sequences. Using shared matches, the unmatched regions of proteins are ignored, meaning similar discrimination can only be achieved for proteins with high coverage.

Significant matches to at least three distinct sequences are required for a family to have two members, (one nonshared match for each member plus one shared match to connect them). Thus, multimember families are more common toward the top of the report, where coverage is relatively high for at least one protein in each family. The relationship between clustering and peptide FDR is more complex. If it is possible to get significant scores for matches to peptide sequences that are so short as to occur by chance in multiple, unrelated proteins, these could create both false protein matches and false connections among proteins. The minimum peptide length in Mascot has a default setting of five residues, and it is difficult to get a significant match to very short peptides. For example, the maximum score that a 5-mer can achieve is 49 (GGGGG with perfect mass accuracy, complete y series, no other peaks). Even so, it is advisable to increase the minimum peptide length if the peptide FDR is unusually high and the number of spectra is large compared with the size of the database. As an example of an extreme case, the peptide FDR was set to 15% for a search of 278,000 spectra against SwissProt 57.11 (512,994 sequences). Overclustering produced a family of 1223 members representing 22,450 same-set and subset proteins. Increasing the minimum peptide length from five to six removed many false connections, but the largest family still had 221 members. Further increasing the minimum peptide length to seven eliminated overclustering, reducing the largest family to nine members, all myosin heavy chain. It would be better if the length threshold was a function of the peptide FDR and the ratio between the number of spectra and the number of entries in the database, rather than a fixed value.

For peptides of reasonable length, which have negligible chance of occurring in multiple, unrelated proteins by chance, false peptide matches can lead to false proteins, but they can only cause false connections among true proteins when there are two or more significant matches to a single spectrum with identical scores for unrelated peptide sequences. The rule that matches with the same score are treated as indistinguishable sequences, which is useful for masking I/L and Q/K interchange, makes false connections possible, and they would not occur if only a single match per spectrum was allowed, however high the peptide FDR. (If the peptide sequence occurs in both proteins, the connection is legitimate even though the peptide match is not.) Unless the significance threshold is reduced to a level at which totally random matches to poor quality spectra are being accepted, it is rare for a spectrum to get significant matches to two unrelated sequences with identical scores, so that false peptide matches rarely create false connections in practice.

The other possibility is under-clustering, where we have peptide matches to two or more sequences with high homology, but no shared matches. We can estimate the extent of this by looking for members from different families that are assigned to the same UniGene cluster. There are several such cases in the report. When inspected, the sequences involved tend to have relatively low homology, which means that there are few if any shared peptides in the translated protein. For example, hits 28 (BG082332) and 31 (AI020624) both map to UniGene cluster Mm.29110 (Ces1f Carboxylesterase 1F). Although the nucleic acid sequences have some homology, they do not have a single shared tryptic peptide in common. That is, even with 100% coverage, we would not see any connection between these two entries on the basis of shared peptide matches.

CONCLUSIONS

A new report has been described that seeks to present database search results in a more logical format, facilitating inspection of peptide match data for individual protein assignments. A greedy set cover algorithm is used to create a minimal set of proteins, grouped into families on the basis of shared peptide matches. For families with multiple members, hierarchical clustering is performed, using the scores of non-shared peptide matches as a distance metric. Dendrograms illustrate how family members are related and can be cut to discard members for which there is judged to be insufficient evidence. Family grouping simplifies the top-level report, making it easier to locate proteins of interest in very large data sets, when the great majority of proteins may be of no interest.

‡ To whom correspondence should be addressed: Matrix Science Ltd., 64 Baker Street, London W1U 7GB, UK. Tel.: +44 20 7486 1050; Fax: +44 20 7224 1344; E-mail: jcottrell@matrixscience.com.

REFERENCES

1. Nesvizhskii, A. I., and Aebersold, R. (2005) Interpretation of shotgun proteomic data - The protein inference problem. *Mol. Cell. Proteomics* **4,** 1419–1440
2. Li, N., Wu, S. F., Zhu, Y. P., and Yang, X. M. (2009) The progress of protein quality control methods in shotgun proteomics. *Prog. Biochem. Biophys.* **36,** 668–675
3. Yang, X., Dondeti, V., Dezube, R., Maynard, D. M., Geer, L. Y., Epstein, J., Chen, X., Markey, S. P., and Kowalak, J. A. (2004) DBParser: web-based software for shotgun proteomic data analyses. *J. Proteome Res.* **3,** 1002–1008
4. Slotta, D. J., McFarland, M. A., and Markey, S. P. (2010) MassSieve: Panning MS/MS peptide data for proteins. *Proteomics* **10,** 3035–3039
5. Kristensen, D. B., Brønd, J. C., Nielsen, P. A., Andersen, J. R., Sørensen, O. T., Jørgensen, V., Budin, K., Matthiesen, J., Venø, P., Jespersen, H. M., Ahrens, C. H., Schandorff, S., Ruhoff, P. T., Wisniewski, J. R., Bennett, K. L., and Podtelejnikov, A. V. (2004) Experimental Peptide Identification Repository (EPIR): An integrated peptide-centric platform for validation and mining of tandem mass spectrometry data. *Mol. Cell. Proteomics* **3,** 1023–1038
6. Resing, K. A., Meyer-Arendt, K., Mendoza, A. M., Aveline-Wolf, L. D., Jonscher, K. R., Pierce, K. G., Old, W. M., Cheung, H. T., Russell, S., Wattawa, J. L., Goehle, G. R., Knight, R. D., and Ahn, N. G. (2004) Improving reproducibility and sensitivity in identifying human proteins by shotgun proteomics. *Anal. Chem.* **76,** 3556–3568
7. Tabb, D. L., McDonald, W. H., and Yates, J. R., 3rd (2002) DTASelect and contrast: Tools for assembling and comparing protein identifications from shotgun proteomics. *J. Proteome Res.* **1,** 21–26
8. Stephan, C., Reidegeld, K. A., Hamacher, M., van, Hall, A., Marcus, K., Taylor, C., Jones, P., Müller, M., Apweiler, R., Martens, L., Körting, G., Chamrad, D. C., Thiele, H., Blüggel, M., Parkinson, D., Binz, P. A., Lyall, A., and Meyer, H. E. (2006) Automated reprocessing pipeline for search-

ing heterogeneous mass spectrometric data of the HUPO brain proteome project pilot phase. *Proteomics* **6,** 5015–5029

9. Zhang, B., Chambers, M. C., and Tabb, D. L. (2007) Proteomic parsimony through bipartite graph analysis improves accuracy and transparency. *J. Proteome Res.* **6,** 3549–3557
10. Ma, Z. Q., Dasari, S., Chambers, M. C., Litton, M. D., Sobecki, S. M., Zimmerman, L. J., Halvey, P. J., Schilling, B., Drake, P. M., Gibson, B. W., and Tabb, D. L. (2009) IDPicker 2.0: Improved protein assembly with high discrimination peptide identification Filtering. *J. Proteome Res.* **8,** 3872–3881
11. Weatherly, D. B., Atwood, J. A., 3rd, Minning, T. A., Cavola, C., Tarleton, R. L., and Orlando, R. (2005) A heuristic method for assigning a false-discovery rate for protein identifications from Mascot database search results. *Mol. Cell. Proteomics* **4,** 762–772
12. Moore, R. E., Young, M. K., and Lee, T. D. (2002) Qscore: An algorithm for evaluating SEQUEST database search results. *J. Am. Soc. Mass Spectrom.* **13,** 378–386
13. Kislinger, T., Rahman, K., Radulovic, D., Cox, B., Rossant, J., and Emili, A. (2003) PRISM, a generic large scale proteomic investigation strategy for mammals. *Mol. Cell. Proteomics* **2,** 96–106
14. Nesvizhskii, A. I., Keller, A., Kolker, E., and Aebersold, R. (2003) A statistical model for identifying proteins by tandem mass spectrometry. *Anal. Chem.* **75,** 4646–4658
15. Sadygov, R. G., Liu, H., and Yates, J. R. (2004) Statistical models for protein validation using tandem mass spectral data and protein amino acid sequence databases. *Anal. Chem.* **76,** 1664–1671
16. Feng, J., Naiman, D. Q., and Cooper, B. (2007) Probability model for assessing proteins assembled from peptide sequences inferred from tandem mass spectrometry data. *Anal. Chem.* **79,** 3901–3911
17. Price, T. S., Lucitt, M. B., Wu, W., Austin, D. J., Pizarro, A., Yocum, A. K., Blair, I. A., FitzGerald, G. A., and Grosser, T. (2007) EBP, a program for protein identification using multiple tandem mass spectrometry datasets. *Mol. Cell. Proteomics* **6,** 527–536
18. Shi, J. H., and Wu, F. X. (2009) Protein inference by assembling peptides identified from tandem mass spectra. *Curr. Bioinf.* **4,** 226–233
19. Kuster, B., Schirle, M., Mallick, P., and Aebersold, R. (2005) Scoring proteomes with proteotypic peptide probes. *Nat. Rev. Mol. Cell Biol.* **6,** 577–583
20. Li, Y. F., Arnold, R. J., Li, Y., Radivojac, P., Sheng, Q., and Tang, H. (2009) A Bayesian approach to protein inference problem in shotgun proteomics. *J. Comput. Biol.* **16,** 1183–1193
21. Li, J., Zimmerman, L. J., Park, B. H., Tabb, D. L., Liebler, D. C., and Zhang, B. (2009). Network-assisted protein identification and data interpretation in shotgun proteomics. *Mol. Syst. Biol.* 5, http://dx.doi.org/10.1038/msb.2009.54
22. Gupta, N., and Pevzner, P. A. (2009) False discovery rates of protein identifications: a strike against the two-peptide rule. *J. Proteome Res.* **8,** 4173–4181
23. Grobei, M. A., Qeli, E., Brunner, E., Rehrauer, H., Zhang, R., Roschitzki, B., Basler, K., Ahrens, C. H., and Grossniklaus, U. (2009) Deterministic protein inference for shotgun proteomics data provides new insights into Arabidopsis pollen development and function. *Genome Res.* **19,** 1786–1800
24. Qeli, E., and Ahrens, C. H. (2010) PeptideClassifier for protein inference and targeted quantitative proteomics. *Nat. Biotechnol.* **28,** 647–650
25. Reiter, L., Claassen, M., Schrimpf, S. P., Jovanovic, M., Schmidt, A., Buhmann, J. M., Hengartner, M. O., and Aebersold, R. (2009) Protein identification false discovery rates for very large proteomics data sets generated by tandem mass spectrometry. *Mol. Cell. Proteomics* **8,** 2405–2417
26. Cochrane, G. R., and Galperin, M. Y. (2010) The 2010 Nucleic Acids Research database issue and online database collection: a community of data resources. *Nucleic Acids Res.* **38,** D1–D4
27. Duncan, M. W., Aebersold, R., and Caprioli, R. M. (2010) The pros and cons of peptide-centric proteomics. *Nat. Biotechnol.* **28,** 659–664
28. Sayers, E. W., Barrett, T., Benson, D. A., Bolton, E., Bryant, S. H., Canese, K., Chetvernin, V., Church, D. M., Dicuccio, M., Federhen, S., Feolo, M., Geer, L. Y., Helmberg, W., Kapustin, Y., Landsman, D., Lipman, D. J., Lu, Z., Madden, T. L., Madej, T., Maglott, D. R., Marchler-Bauer, A., Miller, V., Mizrachi, I., Ostell, J., Panchenko, A., Pruitt, K. D., Schuler, G. D., Sequeira, E., Sherry, S. T., Shumway, M., Sirotkin, K., Slotta, D., Souvorov, A., Starchenko, G., Tatusova, T. A., Wagner, L., Wang, Y., John, Wilbur, W., Yaschenko, E., and Ye, J. (2010) Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res.* **38,** D5–16
29. Alm, R., Johansson, P., Hjerno, K., Emanuelsson, C., Ringnér, M., and Häkkinen, J. (2006) Detection and identification of protein isoforms using cluster analysis of MALDI-MS mass spectra. *J. Proteome Res.* **5,** 785–792
30. Seymour, S. L., Lane, W. S., Nesvizhskii, A. I., Searle, B., Tabb, D. L., and Kowalak, J. A. (2008) RG11 ABRF iPRG2008 Study: assessing the quality and consistency of protein reporting on a common dataset. *J. Biomol. Tech.* **19,** 88–93
31. Searle, B. C. (2010) Scaffold: A bioinformatic tool for validating MS/MS-based proteomic studies. *Proteomics* **10,** 1265–1269
32. Seymour, S. L. (2010) Assessing and interpreting protein identifications. *J. Biomol. Tech.* **21,** S12
33. Ishihama, Y., Oda, Y., Tabata, T., Sato, T., Nagasu, T., Rappsilber, J., and Mann, M. (2005) Exponentially modified protein abundance index (emPAI) for estimation of absolute protein amount in proteomics by the number of sequenced peptides per protein. Mol. & Cell. *Proteomics* **4,** 1265–1272
34. Larkin, M. A., Blackshields, G., Brown, N. P., Chenna, R., McGettigan, P. A., McWilliam, H., Valentin, F., Wallace, I. M., Wilm, A., Lopez, R., Thompson, J. D., Gibson, T. J., and Higgins, D. G. (2007) ClustalW and ClustalX version 2. *Bioinformatics* **23,** 2947–2948
35. Cormen, T. H., Leiserson, C. E., Rivest, R. L., and Stein, C. (2001) *Introduction to Algorithms,* 2nd Ed., MIT Press and McGraw-Hill, Cambridge, MA